(Paper's Draft) The Effects of an Aging Population on Military Recruitment and Retention: A Data-Driven Analysis

Lázaro Damasceno

January 2, 2026

Abstract

Resumo em PT-BR.

Abstract

Resumo em EN.

1 Introduction

2 Theoretical Framework

3 Methodology

3.1 Programming Language For Data Analysis

Python was selected as the programming language for data analysis. The libraries used included *polars*, *dask*, *matplotlib*, *scikit-learn*, and *statsmodels*, distributed as follows:

Table 1: Selected Python Libraries

Library	Used for	Used module
Polars	Data analysis.	None
Dask	Pythonic out-of-memory data analysis.	dataframe

Continued on next page

Table 1 – continued

Library	Used for	Used module
Matplotlib	Plotting figures.	pyplot
Scikit-learn	AI/ML.	preprocessing.StandardScaler, preprocessing.RobustScaler, preprocessing.OneHotEncoder, compose.make_column_transformer, cluster.KMeans or cluster.DBSCAN
Statsmodels	Estimating statistical models, conducting tests, and exploring data.	tsa.arima.model.ARIMA

To ensure reproducibility and avoid conflicts with system-wide settings, Miniconda was used to manage the development environment.

Within Miniconda, a dedicated environment named `etl_eda_ml` was created. All core packages were installed through the Conda package manager. Additionally, to enable Excel, CSV, and ODS file support in *polars*, the *fastexcel* engine was installed via the `conda-forge` channel.

The analysis were implemented in Jupyter notebooks with the specified Conda environment.

When Polars encountered bottlenecks, Apache Spark was used to select variables by dropping unnecessary ones and transforming the remaining data in parquet. This approach reduces the total data volume, allowing Polars to load the remaining data much faster.

3.2 Datasets

The selected datasets were the *Predicted Populations of Brazil (2000-2070)*, made and shared by the Brazilian Census Office (BCO)[1], and the *Military Servive*, made and shared by the Brazilian Army (BA).[2]

The Government data is published on the Brazilian Government Open Data Portal. The portal was institutionalized by the Presidential Decree no. 8.777/2016.[3]

The military data includes 18 CSV files, whose total data volume is 4,9 GB. As Polars was slow to load them, Apacha Spark was used. The total number of columns was 22. 13 were selected. The reduction of the number of columns was, more or less, 40,91%.

[1] https://www.ibge.gov.br/estatisticas/sociais/populacao/9109-projecao-da-populacao.html
[2] https://dados.gov.br/dados/conjuntos-dados/servico-militar
[3] https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm

The next step was turn the military data into parquet. 4,9 GB of data were reduced to 143,6 MB. In other words, the GB data was reduced 97,07%. As the new transformed data is only 2,93% of the original data, Polars can handle it.

4 Discussion

5 Results

6 Final Remarks